UNDERSTANDING OF RELIABILITY, VALIDITY AND PRACTICALITY IN LANGUAGE TESTING

By Nguyen Thi Thu Thuy
Vietnam National University of Agriculture

INTRODUCTION

Language testing is central to language teaching. It provides goals for language teaching, and it monitors, for both teachers and learners, success in reaching those goals. It also provides a methodology for experiment and investigation in both language teaching and language learning/acquisition (Davies, 1990, p.1). Therefore the quality of a language test is very important for both teachers and learners. In this paper, I am going to evaluate the quality of an achievement test planned, designed and conducted in my own university in terms of *reliability*, *validity and practicality*.

The final achievement test is used to test 140 second-year English majored students who have learnt 45 credit hours of English grammar in academic year 2023-2024 at Vietnam National University (VNUA). Their course books are A University Grammar of English students' book and workbook, written by Randolph Quirk & Sidney Greenbaum (1971).

RELIABILITY

According to Lado (1961, p.330), reliability has to do with the stability of the scores for the same individuals. If the scores of the students are stable, the test is reliable; if the scores tend to fluctuate for no apparent reason, the test is unreliable. A test is considered as a reliable test if it measures consistently, i.e., on a reliable test, someone will get more or less the same score, whether they happen to take it on one particular day or on the next, whereas on an unreliable test the score is quite likely to be considerably different, depending on the day on which it is taken (Hughes 1989:3). VNUA grammar's test which was taken by the end of the academic year is reliable, firstly, in the sense that the score obtained by a student is pretty close to the score he would obtain if we gave the test again (Lado, 1961). Secondly, the more items that the test has, the more reliable that test will be (Hughes, 1989, p. 36). The test is long enough to achieve satisfactory reliability as it has 3 questions: Question 1 consists of 10 items for *Divide each of the sentences below into its constituent parts, and label each part S, V, C, O or A*, Question 2 asks students to *Use all kinds of verbs (stative verbs, dynamic verbs, lexical verbs, auxiliary verbs: primary & modal verbs), tenses (present tenses, past tenses, & future forms), mood (indicative mood, imperative mood, and subjunctive mood) and voice (active & passive) to write a short essay (around 190 words) about an important event you and your family have been to.* And Question 3 asks students to *Use all types of determiners (pre, central & post); modification (pre-modification such as 's genitive, -ing/-ed participles, sentence, adverbial, noun & adjective; post-modification such as relative clause, prepositional phrase, non-finite clause, appositive clause, clause of time, place, manner & reason, adverb, adjective; and multiple – modification (pre & post) to create your own story (about 250*

words) based on the 115-word story. Thirdly, the test is reliable because it does not give candidates too much freedom to choose, every candidate has the same task to do. Fourthly, before the test was taken by the candidates, it was given to the head of the department, who is a senior lecturer of English to check whether or not there were any ambigious items; whether the instructions were clear and explicit enough; whether the test was well laidout and perfectly legible and whether the test format and test techniques were similar to those of their progress tests and whether it followed the test specification table. Furthermore, while conducting the actual test, we tried our best to reduce the factors affecting test reliability such as lighting, temperature, distractions or noise; differences in administrative instructions; test compromise (e.g, no students knew the questions and/or answers beforehand); inaccuracy in scoring; inadequate sampling of test items; lack of motivation, fatigue; or illness in the examinees, and improve the reliability of the test by standardizing and optimising the testing conditions; using a uniform procedures in administering the test; increasing the number of the test items; providing an adequate sampling of test items and reducing subjective scoring of the test (Finocchiaro & Sako, 1983. p. 28).

In terms of scorer reliability, 10 items which takes up 20 points out of 100 points are objective and writing 2 compositions (80%) is subjective. To increase scorer reliability, the key was made in details and before the real scoring began, the section which identified the candidates' name, birthday and birth place and their index number was cut off, the code number was written instead. In marking writing compositions, the band descriptors and the marking schemes were subjected to group criticism before the real scoring began. The archetypical representatives of different levels of ability were selected to be marked by all the scorers as samples so as to avoid individuals whose scoring would deviate markedly and inconsistently from the norm. As a general rule, all the scripts were scored by at least two independent scorers. Neither scorer knew how the other had scored a test paper. The scores were recorded on a separate score sheets and passed to a third senior colleague, who compared the two sets of scores and investigated discrepancies (Hughes, 1989).

VALIDITY

Bachman's point of view (1990, p. 289) is that the most important quality in the development, interpretation, and use of language tests is validity, which has been described as a unitary concept related to the adequacy and appropriateness of the way we interpret and use test scores. According to Lado (1961, p. 30), the test is valid if it measures what it is intended to measure. Validity in language depends on the linguistic content of the test and on the situation or technique used to test this content. A test that uses a perfectly valid conversational situation but does not test the elements of the language is not valid. On the other hand, a test that tests the elements of language but does it by lists or rules or techical names rather than in use in essential communicative situations is not a valid test either. Finocchiaro & Sako (1983, p. 24-25) add that two questions must be considered when determining test validity of a foreign language test: What aspects of the language is the test designed to measure, and how well it, in fact, measures the global skills or the dicrete elements of the language? We have to postulate four primary validity concepts: *content validity, concurrent validity, predictive validity, construct validity and face validity*.

Content validity

One outcome of Finocchiaro & Sako's work (1983, p. 25-27) is that content validity is assured by checking all items in the test to make certain that they correspond to the instructional objectives of the course, whether they are discrete or integrated language skills. Content validity consists of test specification, or ability, and of test facets, and the demonstration that the tasks included on the test are representative of those specified in these domains (Bachman, 1990, p. 289-290). Based on these criteria, the achievement test of VNUA has content validity in the sense that it corresponded and indicated the content and objectives of the course the VNUA second-year English majored students pursued. The achievement test was designed to measure students' achievements in grammar subject which is compulsory for English majors in their second year learning at VNUA. The contents of the test presented the representatives of the contents of their syllabus.

Validity, according to Lado (!961), can be achieved and verified indirectly by correlating the scores on a test with those of another test or criterion which is valid. The grammar's achievement test is valid in sense that the two sets of scores (the progress test scores and the achievement test scores) correlate quite highly, i.e., the students who made high scores on the progress test also scored high on the achievement test and those who scored low on the progress test also scored low on the achievement test. As a rule in my university, it is the teacher who has to keep the record of the students' progress and achievement test scores so that the teacher will be able to see how much the students have achieved the objectives of the course and also to see the students' gap. The achievement test scores obtained throughout the year give a detailed picture of the relative strengths or weaknesses in student performance at each important stage of the course (Finocchiaro & Sako, 1983). Five students who performed below 40% of the total scores/ the minimum standards had to redo another achievement test. As determinants of the final standing of students programme (reflecting each student's status relative to that of other students, as well as absolute standards of achievement), grades should show the progress made during the course and measure a student's current level of proficiency.

Concurrent Validity

According to Finocchiaro & Sako (1983), our achievement test has some concurrent validity in the sense that it was administered to students in the course for which the test was developed and scores were recorded for each student. These scores were then compared to teachers' rating. The result was that the individual with the highest grades or teachers' ratings scored highest in the achievement test and those with the lowest grades and/or ratings on the progress tests had also been rated lowest by the teachers. Therefore, the achievement test measured what it was designed to measure.

Predictive validity

Predictive validity is used extensively in the validation of language aptitude tests. However, our university's achievement test had predictive validity in the sense that it was determined to test a group of prospective students to see their progress in their language courses and to see if the

students had achieved enough knowledge or objectives of the course so that they will be able to move to a higher course, i.e., intermediate course.

Construct Validity

In Finocchiaro & Sako's point of view (1983), construct validity should be used extensively in the validation of aptitude tests and proficiency tests, but is far less important in the validation of achievement tests. The achievement test of our university had construct validity because it was used to measure the students' achievements by the end of the course. The test was well constructed in form of 3 sections of which instructions and points were clearly stated. 10 items are multiple choice questions, which are objective and easy enough for the examiners to score.

Face validity

According to Finocchiaro & Sako (1983) and McNamara (2000), our achievement test had some face validity in the sense that it met the expectations of those involved in its use such as the educators, administrators, students, and the general public. It has face validity because it is a test intended to measure the students' achievements. In other words, it is said to have face validity because it measured what it was supposed to measure, i.e., students' achievements in grammar, the language use and writing.

Practicality

In Finocchiaro & Sako's point of view (1983), the criteria of practicality normally will be based on such factors as *economy, scorability and administrability*.

Economy

Lado (1961) and Finocchiaro & Sako (1983) share the same point in saying that the test is practical and economical if it measures what we want it to test in a reasonable time considering the testing situation. As a rule, the achievement test was conducted by the end of the course according to the university's timetable which was given to the students at the beginning of the course. To be economical, the university designed a series of test items for achievement tests which correspond with the table of test specification, which is called 'test bank'. At the examination, each student was given a test booklet and they are allowed to write the answers in their paper. The test is a combination of both objective and subjective test items so the marking is not totally objective. The booklets cannot be reused, which is not economical.

Scorability

In Finocchiaro & Sako's point of view (1983), the practicality of a language test is further determined by its ease of scoring. Tests which are difficult to score become a burden for the scorer, and are unduly demanding of personnel, time and resources. Scoring can be done by hand through the use of scoring keys. Subjective tests such as writing in question 2 and 3 presented problems in scoring since they were designed that they required the students to apply knowledge about grammatical categories to write paragraghs/compositions. One rater may – and likely will – differ in his/her rating from another rater. The incomparable advantage of a test graded objectively

resides in the knowledge that different persons scoring the test will consistently arrive at the same score.

Administrability

According to Finocchiaro & Sako (1983), the achievement test of the university had practicality in the sense that the testing instrument was easy to administer to the examinees. In order to contribute to administrability, before the achievement tests for all courses of the university, a training session for test administrators was done to facilitate the operation and save time and effort later on. Test intructions were clear and concise, and yet totally comprehensible and complete. Instructions which are involved, complex, or unclear will require extensive training of the admininstrator in order for him/her to learn how to administer tests satisfactorily. The layout of items in the test booklet also contributed to the facility of test taking.

CONCLUSION

In short, the major qualities to be considered when developing foreign language tests are test *validity, reliability and practicality*. With relation to validity we noted four primary types: *content, concurrent, predictive, and construct*. The criteria for practicality are based on such factors as *economy, scorability, and administrability*. While validity is the most important quality of test use, reliability is a necessary condition for validity, in the sense that test scores that are not reliable cannot provide a basis for valid interpretation and use. In examining reliability we must identify potential sourses of measurement errors and estimate the magnitude effects on test scores. In investigating validity, on the other hand, we examine the extent to which factors other than the language abilities we want to measure affect performance on test scores. As discussed above, the achievement test which was planned, designed and conducted to measure 140 English majored students' achievements at VNUA by the end of the course was valid, reliable and practical. However, to make this test have more validity, reliability and practicality, there are several problems to be considered, for example, the marking schemes for writing compositions need to be designed in more details.

REFERENCES

1. Bachman, F.L. (1990). *Fundamental Considerations in Language Testing*. OUP.
2. Davies, A. (1990). *Principles of Language Testing*. Basil Blackwell Ltd.
3. Hughes, A. (2003). *Testing for Language Teachers* (2nd Ed) CUP.
4. Lado, R. (1961). *Language Testing*. Longman, Green and Co Ltd
5. McNamara, T.F. (2000) *Language Testing*. OUP
6. Quirk, R. & Greenbaum, S. (1971). *A University Grammar of English*. NXB Giao thông vận tải
7. Quirk, R. & Greenbaum, S. (1971). *A University Grammar of English. Workbook*. NXB Giao thông vận tải